Description

EEPROM WITH MULTI-MEMBER FLOATING GATE

TECHNICAL FIELD

The invention relates to semiconductor integrated circuit memory devices and, in particular, to an electrically erasable programmable read only memory (EEPROM) transistor.

BACKGROUND ART

EEPROM devices employ a floating gate that is a conductive electrode, insulated from other electrodes, used to store electrical charge indicative of the state of the device. The manner of reading the charge on the floating gate is well-known. For example, when charge is present on the floating gate, the device may indicate a binary zero and when charge is not present on the floating gate, a binary one is indicated. Charge is applied to the floating gate by a phenomenon known as Fowler-Nordheim tunneling, or more recently, by band-to-band tunneling. The floating gate was spaced above and separated from the substrate by a layer of thin oxide, having a thickness in the range of 40 or fewer Angstroms to about 70 Angstroms. Since this oxide is applied as a horizontal layer, there is no alignment or deposition problem. In a memory array, millions of identical memory devices are aligned in rows and columns for data storage. Each cell must be an efficient charge storage device. As devices become smaller, it becomes more difficult to maintain charge storage, leading to data retention integrity issues.

In U.S. Pat. No. 5,618,742 to F. Shone et al. a flash EPROM is shown wherein a main floating gate poly body is used for self-aligned placement of source and drain regions. Lateral poly spacers contact the main

poly body to form an extended floating gate. U.S. Pat. No. 6,043,530 to M.B. Chiang, U.S. Pat. No. 6,074,914 to S. Ogura, and U.S. Pat. No. 6,124,170 to M. Lim et al. also show lateral poly spacers.

In U.S. Pat. No. 6,479,351, entitled "Method of Fabricating a Self-Aligned Non-Volatile Memory Cell", assigned to the assignee of the present invention, B. Lojek discloses a self-aligned non-volatile memory cell behaving like an EEPROM with a small, upright, conductive sidewall spacer electrically coupled and being located next to a main floating gate region. Both the small sidewall spacer and the main floating gate region are formed on a substrate and both form the floating gate of the non-volatile memory cell. Both are isolated electrically from the substrate by an oxide layer that is thinner between the small sidewall spacer and the substrate; and is thicker between the main floating gate region and the substrate.

The sidewall spacer has an adjacent thin oxide layer separating it from the main floating region creating a thin dielectric pathway for electrons to tunnel into either the spacer or the main polysilicon floating gate region, both forming the floating gate electrode. Part of the thin oxide layer is deposited vertically between the upright spacer and the main polysilicon body and part of the layer is deposited horizontally. Since the floating gate electrode is maintained at a single voltage indicating a single charge state, the multiple regions of the floating gate must be joined together by a conductive member in order to be at the same electrical potential. A layer of polysilicon applied after conductive spacer formation is usually used to join the floating spacers to the main polysilicon body. This multi-member floating gate construction is

effective for providing multiple paths for tunneling through the thin oxide layer.

Separating the floating gate from the substrate is a layer of thin oxide, perhaps 30-60 Angstroms in thickness. As devices become smaller, the layer of thin oxide becomes more difficult to situate between the substrate and the floating gate. Also, the spacing and position of source and drain electrodes become more difficult to control.

An object of the invention is to provide an EEPROM device having an improved floating gate construction, with improved alignment of tunnel oxide and source and drain electrodes.

Another object of the invention is to provide a compact memory cell geometry with good charge storage and programming efficiency.

SUMMARY OF INVENTION

The above object has been met with a multi-member floating gate construction in a first active area wherein a self-alignment technique allows a uniform deposition of a thin, tunnel oxide layer in an EEPROM device that includes a vertical portion between the main floating gate body and adjacent conductive, upright conductive, sidewall spacers forming part of the floating gate. Prior to thin oxide deposition, the main floating gate body can be used as a mask for ion implantation of source and drain electrodes on opposite sides of the body. The ions are not driven deep, but just barely penetrate the substrate. The drain electrode is much larger than the source electrode, extending into an implanted region that reaches a nearby auxiliary active area, providing an unusually large charge reservoir for programming. The first active area and the auxiliary active area are adjacent, but insulated from each other.

However, they share a large subsurface implanted or doped region which will serve as a reservoir for charged particles to be supplied to the floating gate. The entire footprint of the auxiliary active area is dedicated to this purpose.

A uniform thin oxide layer is deposited in a self aligned manner by first forming the main polysilicon body over an insulating layer of gate oxide and then etching back the gate oxide to the polysilicon body. It is very important that the tunnel oxide layer be both thin, preferably about 10 Angstroms, and uniform. This step is followed by vapor deposition of the thin tunnel oxide layer over the polysilicon body including its side walls. A subsequent layer of polysilicon is applied over the thin oxide layer and etched away, but not totally, leaving small corners of polysilicon adjacent to and against the main polysilicon body in a circumferential loop but separated therefrom by the previously deposited thin oxide. The residual polysilicon appears similar to nitride sidewall spacers, except that the newly formed spacers are conductive polysilicon spacers that are electrically joined to the main polysilicon body in a subsequent step.

The thin oxide between the poly spacers and the main polysilicon body allows electrons to have bidirectional tunneling opportunities, either into the main polysilicon body or into the poly spacers. This bi-directional tunneling opportunity increases the probability of electron transfers to and from the floating gate in charge transfer operations, such as write or erase operations, particularly using low voltages, and allows use of smaller polysilicon structures in forming the floating gate.

An insulative layer is deposited over the polysilicon members and later a conductive cap is

deposited over the insulative layer joining multiple members of the floating gate. A hole formed through the cap is filled with metal, allowing joinder of the conductive cap to underlying floating gate members so that the main polysilicon body and the conductive spacers are at the same electrical potential. The extended drain structure, cooperating with the multi-element floating gate structure leads to a compact memory cell geometry with good charge storage and programming efficiency.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a top view of first and second active areas defining memory cell footprints on a semiconductor substrate in accordance with the present invention.

Fig. 2 is a side sectional view of the memory cell footprint illustrated in Fig. 1, taken along lines 2-2, with isolation regions bordering active areas.

Fig. 3 is a top view of a memory cell footprint as in Fig. 1, with an overlay of a mask for doping substrate regions.

Fig. 4 is a side sectional view along lines 4-4 in Fig. 3.

Fig. 5 is a top view of a memory cell footprint as in Fig. 3, with a first polysilicon layer etched to a rectangular pattern.

Fig. 6 is a side section view along lines 6-6 in Fig. 5.

Fig. 7 is a top view of a memory cell footprint as in Fig. 5, with a tunnel oxide layer etched to a rectangular pattern and an overlying second polysilicon layer overlying the oxide layer removed except for a poly spacer region appearing as a rectangle.

Fig. 8 is a side sectional view along lines 8-8 in Fig. 7.

Fig. 9 is a top view of a memory cell footprint as in Fig. 7, with a third polysilicon layer overlying other layers.

Fig. 10 is a side sectional view along lines 10-10 in Fig. 9, showing source and drain implants and a TEOS layer, all applied before the third polysilicon layer of Fig. 9.

Fig. 11 is a side sectional view showing the structure of Fig. 10 with a photomask over the central raised third polysilicon layer.

Fig. 12 is a top view of a memory cell footprint as in Fig. 9, with a new third polysilicon boundary formed by etching to the substrate.

Fig. 13 is a side sectional view along lines 13-13 in Fig. 12 with additional ILD and contact mask layers over the entire structure.

Fig. 14 is a side sectional detail of a cut through the layers shown in Fig. 13.

Fig. 15 is a top view of a memory cell footprint as in Fig. 12, with contact positions shown.

Fig. 16 is a side sectional detail as in Fig. 14, with the cut being filled and capped.

DESCRIPTION OF THE INVENTION

With reference to Fig. 1, a footprint 21 defines a first active area of a memory transistor. A second footprint 23, smaller than footprint 21, defines a second active area of an auxiliary region to be used to apply a tunneling or programming voltage. The entire auxiliary region 23 will become part of a subsurface doped region, extending to a stripe crossing the width of the first active region 21, and having a sufficient ion reservoir for providing sufficient charge to the floating gate for programming. The active areas are those silicon wafer areas that are doped and used in transistor

construction. The regions outside of the active areas are isolated, as will be seen below. For example, lateral zones 22, 24, and 26 are outside of the active area and will be isolated so that one active area does not communicate with another, except by shared electrical paths. The implanted auxiliary region 23 may exceed the area of the shared stripe path, an implanted region, crossing the width of the first active area 21. Regions 28 and 29 are width-wise zones which are outside the active area and similarly isolated. The central region defined by footprint 21 is the main active area 25. An extension 27 to the left of the main active area provides for a source electrode of different and much smaller length dimension than the drain, as will be seen below.

Turning to Fig. 2, doped substrate 31 is a semiconductor body having an upper surface 33 upon which features are constructed. Features may also be constructed by implantation into the substrate 31 as will be discussed below. Isolation regions 32, 34, and 36 are shallow trench isolation (STI) regions which separate one active area of the substrate from another.

After the active areas have been defined, a layer of screen oxide 43, shown in Fig. 4, i.e. a thermal oxide, is grown over the entire surface prior to positioning the implant mask 41, shown in Fig. 3. The implant mask defines a N+ region where ions will be implanted, for example, arsenic ions. The implant region defines a wide stripe across the main active area 25 which will serve as a supply or sink for electrons to be transferred to the floating gate structure. The stripe across the entire width of the active region 25 ensures a good supply of charged particles, while allowing some slight misalignments in positioning the floating gate. The stripe is enhanced by implanting over the entire auxiliary active region 23, or at least most of it. The

auxiliary region assures a good supply of charge for movement to the charge storage regions, namely the floating gate.

In Fig. 4, it is seen that the screen oxide layer 43 has been stripped off of the isolation regions 28 and 29. The ion implant region 45 is a N+ subsurface highly doped region in the substrate 31 immediately below the upper surface 33, occupying the implant mask region 41.

In Figs. 5 and 6, those portions of the gate oxide layer extending beyond the footprint of the first etched poly layer 51 are removed. Upon removal, there is a single ion implantation of electrodes including a first electrode implant region 71 and a second electrode implant region 73, with the first electrode implant region being a future source electrode and the second electrode implant region 73 being a future drain electrode implant. Note that the implant regions are barely below the surface of the substrate. The main floating body 51 acts as alignment mask for the first and second electrode implant regions 71 and 73 so that these electrodes are essentially self aligned with the left and right edges of the rectangular main floating body 51.

Once the implantation is complete, the screen oxide may be stripped from exposed areas. However, a layer of gate oxide 53 remains below the main floating body 51. Gate oxide 53 is a rectangular deposition covering most of the main active area 25, as well as a portion of the footprint 23 of the auxiliary active area. Atop the gate oxide layer is first polysilicon layer 51 etched to a rectangle mostly over the main active area 25. The polysilicon layer 51 is etched to the rectangular shaded pattern seen in Fig. 5. Over the top of the first polysilicon layer 51, a very thin tunnel

oxide layer 61 is grown which may be more clearly seen in Fig. 8.

In Fig. 8 and as described above, the gate oxide layer 53 is seen to be above substrate 31. The first polysilicon layer 51, etched to a rectangular main floating body, is seen to be above the gate oxide layer. Now, the thin tunnel oxide layer 61 is grown over the main floating body, extending over the substrate on both sides. The thin oxide layer has a vertical portion 63 rising along a vertical wall of the first polysilicon layer and then having a top mesa portion 65, sometimes called poly oxide. The thickness of the thin oxide layer is in the range of 30-70 Angstroms.

In Fig. 8, conductive polysilicon spacers 67 and 69 are seen to resemble nitride spacers of the prior art and be insulated from the first polysilicon layer 51 by the vertical portion 63 of the tunnel oxide. The spacers are also insulated from the substrate by the tunnel oxide layer 61. It will be seen that both the main floating body 51, as well as the poly spacer 69 both reside above the subsurface ion implantation region 55. The ion implantation region can supply charge, i.e. electrons, to both the main floating body 51 as well as the poly spacer 69 through a vertical portion 63 of the tunnel oxide. In other words, electrons migrating through the tunnel oxide have two paths to reach a floating member, with one path being into the main floating body 51 and the other path being into the poly spacer 69. This is thought to enhance the probability of electron capture into a floating body.

With reference to Fig. 9, a third poly layer 77 is deposited over the top of the structure seen in Fig. 8. The function of the third poly layer will be to electrically connect the first and second layers. In Fig. 10, the third capping dielectric layer 77 is seen to

itself be capped by a dielectric layer 75 which may be a TEOS layer which follows the contours of the structure below so that the mesa feature described above is preserved. In Fig. 11, the mesa feature has been simplified by a single body 83 which includes all of the elements seen in Fig. 10, both above the upper surface of the substrate and below. A photomask 81 is applied over the mesa feature, well within the isolation regions 28 and 29. The photomask 81 allows removal of all layers to the outside of the mask, stopping at the substrate. Once all layers outside the mesa structure are removed, the mask itself is removed.

Considering Figs. 12, 13, and 14, an interlayer dielectric (ILD) region 85 is applied directly over the substrate and over the mesa feature. A photoresist contact mask 87 is applied above the dielectric layer 85. A cell contact 89 is located longitudinally outside of the active area over a portion of the isolation region. In Fig. 13 it may be seen that a hole 82 is patterned in the photoresist layer 87, with the hole extending into the mesa feature, with the hole near but not contacting poly spacer 69. The hole extends through a portion of the rectangular main floating body 51. In Fig. 14 the hole may be seen to extend through the floating body 51, the poly oxide 65, the third poly layer 77, the ILD layer 85, and the contact photomask layer 87.

In Fig. 15, the hole 82 is seen to be filled with a metal filler 91, preferably a tungsten plug which may be sputtered into the hole. Poly spacer 69 is seen to make contact with third poly layer 77. In turn, the metal plug 91 connects the third layer 77 with the main floating body 51 so that the main floating body and the poly spacer are at the same electrical potential, a very important connection.

In Fig. 16, it will be seen that the contact mask may be provided with additional contacts 80, 82, and 84, 86 and 88 for making electrical contact with outside connections. The contacts 84, 86 and 88 are closely spaced along the implant region 41 to spread current uniformly. In Fig. 17, the first implant has been driven to form source electrode 60 and the second implant has been driven to form drain electrode 70. An inherent capacitance C_{gs}, 92, exists between the source electrode 60 and the floating gate main body 51. A second inherent capacitance C_{gc}, 94, exists between gate 51 and substrate 31. A third inherent capacitance C_{gd}, 96, exists between the main body 51 and drain 70.

Returning to Fig. 16, the dimension L_s indicates the relative length of the source, while the dimension L_D indicates the relative length of the drain. L_D is much greater than L_s. This leads to C_{gs} being much smaller than C_{gd}.

This allows a high capacitance coupling of the floating gate voltage during programming which will be approximately

$$V_{FG} \approx \frac{C_{gd}}{\sum C} V_{plate}$$

where V_{plate} is the voltage applied to the subsurface implant region 41 and where $\Sigma C = C_{gs} + C_{gd} + C_{gc}$. Considering applied voltages V_g may be calculated as follows:

$$V_g = \frac{C_{gs}}{\sum C} V_S + \frac{C_{gd}}{\sum C} V_D + \frac{C_{gc}}{\sum C} V_B$$

where V_s is the voltage applied to the source electrode, V_D is the voltage applied to the drain electrode and V_B is the substrate bias voltage. The mechanism for charge

transfer is Fowler-Nordheim or band-to-band tunneling and the reading of charge is accomplished in the usual way.

Typical programming voltage will be on the order of 5 volts, thereby allowing low voltage programming. Other voltages will be less than 5 volts, allowing memory arrays having memory devices as described herein to efficiently store charge in compact memory cells as described herein.